

Wikipediaのカテゴリ階層関係の分類を用いた 日本語 Wikipedia オントロジーの分析

Analysis of Japanese Wikipedia ontology based on the Wikipedia Category Structure Analysis

吉岡 真治*1

Masaharu YOSHIOKA

*1 北海道大学大学院 情報科学研究科

Graduate School of Information Science and Technology, Hokkaido University

Wikipedia is a free encyclopedia on the Internet and contains large numbers of articles. In order to utilize semantic information in this encyclopedia, Wikipedia ontologies (e.g., YAGO2, Japanese Wikipedia ontology) are constructed based on its information. In these ontologies, category information of each article are used for constructing concept hierarchies. However, since Wikipedia category is designed for navigating users to find appropriate article, it is not equivalent to concept hierarchy for Wikipedia ontology. In this paper, we analyze category structure of Japanese Wikipedia and compare this results with class hierarchy of Japanese Wikipedia ontology.

1. はじめに

Wikipedia には、現在、日本語版^{*1}に 898,433 件、英語版^{*2}に 4,465,757 件の記事^{*3}があり、これらのページが階層的なカテゴリによって、分類されている。この階層的なカテゴリは、主に、ユーザーのナビゲーションを支援するものであるが、そのカテゴリ中に、意味的な階層関係を多く含むことから、この階層的なカテゴリの性質を利用した Wikipedia オントロジーの構築が行われている。ただし、実際の Wikipedia に含まれるカテゴリ構造については意味的な階層関係以外の関係を多く含むことから、ヒューリスティクスなどに基づいた関係の抽出などが行われてきた。

本研究では、これまでに、Wikipedia のカテゴリ階層の分析 [藤原 12] や、Web 上に公開された多くの有用なコンテンツを活用するために、複数のコンテンツ間の関係 (Linked Data と呼ぶ) を記述して利用する Linked Open Data (LOD) [Bizer 09] の中心として、Wikipedia を利用するための問題点の議論を行ってきた [吉岡 12]。

本稿では、これまでの研究を踏まえ、Wikipedia のページやカテゴリが持つ性質について網羅的な分析を行うと共に、日本語オントロジーにおけるクラス階層関係との対応関係について議論する。

2. Wikipedia のカテゴリと日本語 Wikipedia オントロジー

2.1 日本語版と英語版 Wikipedia におけるカテゴリの定義

ここでは、日本語版 Wikipedia のカテゴリに関する説明ページ「Wikipedia:カテゴリ」、「Wikipedia:カテゴリの方針」をもとに、カテゴリがもつ性質について概観する [吉岡 12]。

Wikipedia におけるカテゴリの定義は、「カテゴリとは、記事を分野別にまとめた索引」である。これらのカテゴリは、「総記」、「学問」、「技術」、「自然」、「社会」、「地理」、「人間」、「文化」、「歴史」の 9 個の主要カテゴリのいずれかに分類される。

連絡先: 吉岡真治, 北海道大学大学院情報科学研究科, 札幌市北区北 14 条西 9 丁目, 011-706-7107, yoshioka@ist.hokudai.ac.jp

*1 <http://ja.wikipedia.org>

*2 <http://en.wikipedia.org>

*3 2014 年 3 月 6 日現在

さらに、「Wikipedia:カテゴリの方針」においては、カテゴリは図 1 に示すように定義されている。また、カテゴリの階層構造には、図 2 に示すような関係を満たすことが求められる。

また、カテゴリに包含する記事数が増えた場合には、より具体的なサブカテゴリが作成される。このような基準で作成されたサブカテゴリには、上位カテゴリの内容をおおむねカバーしているようなサブカテゴリの組 (例えば、「アジアのサッカークラブ」に対して、「日本のサッカークラブ」、「シンガポールのサッカークラブ」など) が作成できる場合と、「映画作品」に対する「アカデミー受賞作品」の様に、「アカデミーを受賞しなかった映画作品」といった意味のないサブカテゴリを作らないとこの様な組が出来ない場合がある。ただし、後者の場合には、意味のないカテゴリを作らないため、サブカテゴリの網羅性が保証されない。

この前者のように、上位カテゴリの内容をおおむねカバーしているようなサブカテゴリの組を、「分割として機能するカテゴリ」と呼び、後者を「分割として機能しないカテゴリ」と呼ぶ。また、「分割として機能するカテゴリ」については、ページに親カテゴリを重複して付与しないことになっている。

これに対し、英語版の Wikipedia のカテゴリの説明である「Wikipedia:Categorization#Category_tree_organization」においては、カテゴリを、2 つの異なるタイプのカテゴリの存在とその組み合わせという形での整理を行っている (図 3)。

具体的には、話題 (topic) を表すようなカテゴリ (例: フランス) と集合 (set) を表すようなカテゴリ (例: 都市) の 2 種類のカテゴリを考え、前者のカテゴリとページの関係は、トピックの関連性 (図 1 の 2) に相当するものが多いのに対し、後者は意味的階層に相当するもので、カテゴリ間の階層関係や、カテゴリとページの関係は、クラス-サブクラスの関係や、クラス-インスタンスの関係などを含む事が示唆されている。

また、カテゴリには、これらの 2 つのタイプのカテゴリに加え、それらの組み合わせとしての集合-話題 (set-and-topic) のカテゴリ (例、フランスの都市) が存在すると説明されている。この集合-話題のトピックは、日本語版 Wikipedia の説明における記事数が多くなった際に作成されるカテゴリに相当すると考えられる。また、この様にして作成されたカテゴリは、基本的に、そのカテゴリが持つ集合、話題の両者を上位カテゴリに持つ。例えば、「フランスの都市」の場合は、「フランス」

1. カテゴリは第一義として、「分類」を示すものです。「xx は YY のひとつである」と言うことができれば、「分類」を示すと言えます。項目 xx はカテゴリ YY に属するべきです。反例として、北朝鮮と韓国は関連がありますが、どちらかがどちらかを包含する関係ではありません。
2. 上記に加えて、ウィキペディアのカテゴリとしては「関連が深いキーワード」を示すことができます。「分類」より「キーワード」を指向しているカテゴリも存在します。記事 xx が「YY 関連用語」であるという意味合いでカテゴリ YY に属することが期待される場合があります。例として、学術用語と Category:学問の関係など。この場合も、カテゴリはより上位の概念であることが求められるため、逆の関係ではありません。
3. また、カテゴリはウィキペディアの骨組みの意味を持ちます。

カテゴリ機能の普及によって、カテゴリの構造がウィキペディアの全体構造を示すこととなりました。カテゴリ同士の関係もウィキペディア全体を意識した一貫性や無矛盾性が求められ、よいカテゴリ構造を作ることが、わかりやすいウィキペディアを作ることにつながります。似た意味合いのカテゴリや大きく重複するカテゴリがある場合は、なるべく内容をすり合わせ、統合を検討しましょう。併存させる場合も、明確な使い分けの方針を決めましょう。そうしなければ混乱が永続することになります(例:「文房具」と「事務用品」など)。

図 1: カテゴリの定義

多くのカテゴリは一つ以上の親カテゴリを持ちます。例えば、Category:日本の作家は Category:各国の作家と Category:日本の人物(職業別)の両方に含まれています。あるカテゴリを他のカテゴリのサブカテゴリとする場合、前者のカテゴリの内容が(ある程度の例外はありますが)後者のカテゴリの内容として含まれるものであることを確認してください。カテゴリの上下関係は親子関係であり、ループ構造にならないように注意してください。ある二つのカテゴリ同士に深い関係があり、しかし上下関係を作らないような場合は、カテゴリの本文で関連づけるに留めてください。

図 2: カテゴリの構造

There are following two main kinds of category.

- Topic categories: are named after a topic (usually sharing a name with the Wikipedia article on that topic). For example, Category:France contains articles relating to the topic France.
- Set categories: are named after a class (usually in the plural). For example, Category: Cities in France contains articles whose subjects are cities in France.

Sometimes, for convenience the two types can be combined, to create a set-and-topic category (such as Category:Voivodeships of Poland, which contains articles about particular Voivodeships as well as articles relating to Voivodeships in general)

図 3: 英語版 Wikipedia におけるカテゴリの説明

「フランスの地理」「フランスの都市」というカテゴリ階層と、「都市」「各国の都市」「フランスの都市」というカテゴリ階層を持つ。ただし、そこに属するページやカテゴリは、集合のカテゴリ「都市」の部分集合である事が求められることは明記されている。一方で、話題のカテゴリである「フランス」の部分集合である事については述べられていない。

日本語版 Wikipedia のカテゴリ階層構造についても、英語版の Wikipedia の構造を反映する形で構成されていると考えられるため、この英語版の説明を日本語版 Wikipedia にあてはめて考えることについても一定の妥当性があると考えている。

このことを踏まえると、Wikipedia のカテゴリ階層構造から、一般的な概念階層の情報を抽出するためには、集合(set)のカテゴリに注目する必要があると考えられる。一方、話題(topic)のカテゴリの階層については、分割のためのある種の属性に関する制約を与えている場合が多いと考えられることから、属性を考慮したインスタンスレベルの類似性の判定などには有用な情報となると考えられる。

2.2 日本語 Wikipedia のカテゴリ階層の分析

我々は、2012年2月6日のダンプデータに存在した100,997件から「スタブ」などの Wikipedia 固有のカテゴリを除去した95,765件のカテゴリとそのカテゴリ間の関係203,975ペアについて、その表記パターンと階層関係についての分析を行った[藤原 12]。

この研究では、Wikipedia のカテゴリを構成する要素に基本的なパターンがあることに注目した分類を行った。この研究では、一般的な、「名詞」、「助詞」、「動詞句」、「接続詞」という品詞分類に加え、「修飾節(例えば、『かつて存在した』)」と「付加情報(例えば、曖昧性回避のための文末の()表記『(業種別)』)」の組み合わせでカテゴリの分類を行った。例えば、「かつて存在した日本の企業(業種別)」というカテゴリは、「かつて存在した{修飾節}+日本{名詞}+の{助詞}企業{名詞}{(業種別){付加情報}」の組み合わせと判断される。

ここで、「修飾節」と「付加情報」は、概念の階層構造を分析するためのパターンとしては、あまり有用でないと考え、これらの項目を無視して、パターンの数を数えたところ、表1のような件数となった*4。

表 1: Wikipedia のカテゴリの表記パターンによる分類

カテゴリのパターン	件数
名詞単独: 「日本」、「宇多田ヒカル」	42,044
名詞+の+名詞: 「日本の野球選手」	50,643
名詞+の+名詞+の+名詞: 「京都市の寺院の画像」	1,141
名詞+に+動詞句+名詞: 「商業に関する学科」	785
名詞+を+動詞句+名詞: 「鉄道を題材にした作品」	706
その他	446

ここで、先ほどの、カテゴリに関する話題、集合の分類を用いて考えると、複数の名詞を含むカテゴリの記述においては、多くの場合において、『「話題(フランス)」の「集合(都市)」』といった形で、後半の名詞が集合を表す場合が多いことが想定される。

また、実際に、カテゴリ間の親子のペアを分析したところ、「名詞単独」「名詞 A + の + 名詞 B」の形式では、27,356件中、親の名詞と名詞 A が同じ(「日本」「日本の人物」)ペ

*4 表の値は、論文執筆後に再検討を行った結果を反映しているため、[藤原 12]とは異なる。

アが 14,051 件、名詞 B が同じ (「作家」「日本の作家」) ペアが 5,378 件と大部分を占めた。これは、「名詞 A + の + 名詞 B」が先に述べた「名詞 (話題) + の + 名詞 (集合)」という形式が多いという想定とも合致する。

また、「名詞 A + の + 名詞 B」「名詞単独」の関係では、名詞 B と子の名詞の間の関係が強く、「日本の歌」「演歌」のようなクラス・サブクラスの関係や、「大阪府の大学」「大阪大学」のようなクラス・インスタンスの関係が存在した。これは、名詞 B に相当する名詞が集合の名詞であるという判断とも一致する。

2.3 日本語 Wikipedia オントロジーにおけるカテゴリ階層の利用

Wikipedia に記述されている様々な概念やインスタンスの情報を活用するために、Wikipedia オントロジーの構築が行われている。代表的なものとしては、英語版 Wikipedia に基づいた YAGO2 (Yet Another Great Ontology 2) [Hoffart 13] や、日本語 Wikipedia に基づいた日本語 Wikipedia オントロジー [玉川 10] が存在する。

これらの研究では、Wikipedia のカテゴリが持つ図 2 の 1 の分類としての役割に注目し、カテゴリ情報からクラスの情報を生成し、そのカテゴリに属するページをインスタンスとして分類するという形で、多くのインスタンスを含む大規模オントロジーの構築を行っている。

しかし、これらのオントロジーでは、前節で述べたカテゴリに関する議論とは異なり、集合-話題のカテゴリについても、クラス階層の一部に含む形となっている。

この内、日本語 Wikipedia オントロジーに注目する。このオントロジーでは、Wikipedia のカテゴリ階層から文字列照合のヒューリスティックスを用いることにより is-a 関係を構築している [玉川 10]。具体的には、「空港」「日本の空港」というように、後方の文字列が一緒の場合に、「空港」「日本の空港」という is-a の関係を抽出する方法と、「日本のスポーツ選手」「日本のゴルファー」というように、前方の文字列が一緒の場合に、その一致した文字列を除去した「スポーツ選手」「ゴルファー」という関係を抽出する方法である。

これらのヒューリスティックスを、集合-話題のカテゴリの多くが「名詞 (話題) + の + 名詞 (集合)」の形で記述されることを踏まえて考察する。前者については、集合-話題のカテゴリの説明にあるように、集合の包含関係が期待される関係であるため、「日本の空港」をクラスとして認めることが適切か否かという議論は別にすると、適切な階層関係を抽出する可能性が高いことが確認できる。また、後者については、「名詞 (話題) + の + 名詞 (集合)」において、話題が共通の場合と考えると、その集合の間には、包含関係が期待されることから、同じく適切な階層関係を抽出する可能性が高いと考えられる。また、単純な「名詞」「名詞」には、必ずしも、is-a の関係ではない話題の階層関係を含む可能性があること、多くの集合の階層関係は、分割のための階層として、集合-話題のカテゴリを持つことなどを考慮すると、精度、再現率の両方の観点からも有用な手法であることが確認できる。

3. 日本語 Wikipedia オントロジーの分析

3.1 日本語 Wikipedia カテゴリの階層関係の分類と利用

2.2 節で述べた分析手法を用いて、日本語 Wikipedia のダンプ (2013 年 8 月 18 日版) からカテゴリの名称とその親子関係の分類を行った。このダンプ中のカテゴリの総数は、121,346 件から「スタブ」などの Wikipedia 固有のカテゴリを除去した 99,902 件のカテゴリとそのカテゴリ間の親子関係 208,999

ペアを用いて、以下の分類に基づく概念間の関係を手作業で作成した。

概念階層 「スポーツ」「テニス」などの意味的な包含関係のある語の関係。

クラス-インスタンス 「惑星」「火星」などのように、後者が前者の具体物である関係。

列挙 「88 星座」「いて座」のように、対応する後者を列挙したものが「前者」の名前であるような関係と、「民放ネットワーク」「ANN」などの様に、名前とその構成要素のような関係。

地理的包含関係 「日本」「北海道」などのように、地理的な包含関係のある語の関係。また、各々の地名は地名のインスタンスとしても扱う。

関連語 「経済学」「経済書」などのように、関連性の存在は確認できるが、上記のどの関係でもない関係。

一般に、これだけ多数のペアを手作業で分類することは困難であるが、このペアの多くが、2.2 節で述べたようにな、「名詞 (話題)」や「名詞 (集合)」を共有しているものであることが多く、「A の B」という形の表現 58,193 件で用いられる A の部分の異なりが 18,449 件、B の部分の異なりが 5,525 件とあまり多くない。また、これらの多くは、分割として機能するカテゴリであり、抽出可能な名詞間の関係も、ペアの数ほどは多くない。例えば、「各国」と「日本」の関係などは、「各国の都市」「日本の都市」「各国の道路」「日本の道路」など 596 件も存在する。また、分割のためのカテゴリでは、その性質上、「各国の都市」「イギリスの都市」といったように、類似の関係を持つ名詞を容易に収集可能である。

「各国の潜水艦」「海上自衛隊の潜水艦」のように、例外的な事象は、存在するが、その頻度は、1 回や 2 回といった少ない頻度であるため、「各国」と対応する語を頻度順にソートしてチェックを行うだけで、容易に「各国」と特定の関係 (この場合は、クラス-インスタンス) を持つ語を収集することができる。

この様な手法を用いて、日本語版 Wikipedia に存在する名詞間の関係について、次の基準で分類を行った。その結果、表 2 に示す関係を抽出した。ここで、クラスの異なり数は 13,132 件であった。

表 2: Wikipedia から抽出した名詞間の関係

概念階層	18,114
クラス-インスタンス	15,648
列挙	1,050
地理的包含関係	5,136
関連語	4,159

これらの語の関係リストを用いて、Wikipedia のカテゴリ階層と概念階層の関係について考察を行う。表 2 にあるように、Wikipedia のカテゴリには、多くのインスタンスの情報があり、これらのインスタンスをカテゴリとするカテゴリ階層からは、概念階層の情報を取り出すことができない。しかし、先に作成したインスタンスのリストを用いてカテゴリの親子関係を調べたところ、インスタンスを親カテゴリに持つ親子関係が 28,214 件、子カテゴリに持つ関係が 25,458 件、後方の文字列を共有しているが、前方の文字列の少なくともどちらかがインスタンス (例:「日本のアルバム」「ゆずのアルバム」) が 41,918 件、子のカテゴリが「名詞 (インスタンス) の名詞 (親カテゴリ)」となっているもの (例:「令」「日本の令」) が 3,859 件、存在した。これらのカテゴリ関係だけでも、全体の

約 48%(99,449 件)であり、日本語 Wikipedia のカテゴリ親子関係には、「名詞(話題)」に相当するインスタンスを用いた分割による関係が中心となっていることが確認できた。

また、作成した概念階層について調べたところ、親子関係がそのまま概念階層となっている関係が 14,622 件であり、3,492 件の関係は、カテゴリ中には、直接存在しないことが確認された。これらの関係は、「名詞(話題)」などの前方の文字列を共有して、後方の文字列が概念階層(例:「BBC の番組」「BBC のニュース番組」)となっている 22,399 件の関係の中から抽出されたこととなる。

また、その他の関連語などの関係なども利用したところ、195,115 件(約 93%)の親子関係については、概念階層を得るのに適さない、もしくは、これまでに作成した概念階層の情報を含むものとなった。残りの 13,884 件については、主に「名詞」「名詞」に関する親子関係や、「NHK のアニメ作品」「NHK アニメの放送枠」といった例外的な共有のパターンを含むために、その判定が難しかったものである。この中から、一定の概念階層の情報を取り出せる可能性はあるが、それほど多くの関係は取り出せないのではないかと考えている。

3.2 日本語 Wikipedia オントロジーのクラスとクラス階層

現在公開されている 2013 年 11 月 7 日のダンプを元にした日本語 Wikipedia オントロジーには、166,937 件のクラスと 58,819 件の is-a 関係が登録されている。

このクラスの内、39,355 件は対応するカテゴリ名が存在するが、127,582 件については、対応するカテゴリ名が存在しない。これらのカテゴリの多くは、カテゴリ以外の情報から作成されたクラス名だと考えている。また、こちらで作成したインスタンスのリストと照合したところ、2,276 件のインスタンスであり、25,291 件については、カテゴリ名中にインスタンスの情報を含む(例:「ゆずのアルバム」)ものであった。また、今回作成したクラスと一致したものは、10,100 件のみであった。本手法で作成したクラスより 3,000 件程度少なく、機械的な抽出を行った際の再現率の問題があることが確認された。

また、58,819 件の内、21,276 件は、子のカテゴリが「名詞(インスタンス)の名詞(親カテゴリ)」となっているものであった。この件数が元々のカテゴリ中の場合に比べて多い理由は、カテゴリ階層の「DJ」「各国の DJ」「フランスの DJ」から、「DJ」「フランスの DJ」という関係を作成しているためと考えられる。また、その他のインスタンスに関連する関係が、7,274 件あり、約 49%(28,550 件)がインスタンスに属する関係であった。それ以外にも、「アニメ作品」「アニメ作品(分類別)」といった曖昧性解消のためのカテゴリが、11,820 件、「A」「A の B」や「B」「A の B」という分割のためのカテゴリに対応するものが 4,748 件存在し、辞書的な意味での一般的な概念階層とは異なるものを多く含んでいることが確認された。

また、クラス階層については、日本語 Wikipedia オントロジーには、本研究で作成した 18,523 件中の 4,338 件しか存在しないため、こちらを利用して、クラス階層関係を充実することが可能であると考えられる。

3.3 考察

本研究では、日本語版 Wikipedia に存在するカテゴリの親子関係の約 93%を用いて、網羅的に階層関係を分析し、その特徴について分析を行った。その結果、概念的な意味階層を含む階層関係は、話題ごとのカテゴリに分割されたあとのカテゴリ間の関係を含めても、約 18%(14,622+22,399 件)程度しか存在しない。よって、日本語 Wikipedia の階層構造を辞書的

な意味での類似性を判定する場合に用いる際には注意が必要である。

ただし、名詞(話題)が与える属性に関する制約に注目し、「北海道出身の人物」に対し、「東京都出身の人物」については、「日本出身の人物」が共通するのに対し、「中国出身の人物」については、共通するカテゴリが「日本出身の人物」より 1 段上位の「出身地別の人物」となるので、「東京都出身の人物」の方が近いと考える事が有用な場合も存在する。

これらを踏まえて、日本語版 Wikipedia に存在するカテゴリの階層関係を利用する場合には、目的に応じて、適切な関係を選択する必要がありと考えられる。

次に、日本語 Wikipedia オントロジーについては、クラスの定義に、辞書的な意味に対応するカテゴリだけでなく、集合-話題の組み合わせで作られているカテゴリを多く含むと共に、辞書的なクラス階層については、不十分な情報しか持っていないことが確認された。

今回の作業によって作成した辞書は、機械的に作成された日本語 Wikipedia オントロジーでは、抽出できていなかった関係の抽出が行えるだけでなく、辞書的な意味での一般的な概念階層とは異なるものを取り除いた概念階層を作成可能であるため、補完的な運用について検討が必要である。

4. まとめ

本稿では、本研究でこれまで行ってきた Wikipedia カテゴリに関する分析手法に基づいて、実際に、日本語版 Wikipedia のカテゴリ階層からの概念階層、クラス-インスタンス関係の抽出を行い、その結果に基づいて、日本語 Wikipedia オントロジーに関する分析を行った。

今後は、この検討を元にした英語版 Wikipedia の分析や日本語 Wikipedia オントロジーの補完方法などについて検討を行っていきたいと考えている。

謝辞

本研究を行うにあたり、日本語 Wikipedia オントロジーの開発者である慶応大学の玉川 奨様には、データ提供ならびに議論に参加して頂いた。また、NII の武田英明先生からもコメントを頂いた。また、本研究の一部は、科研費基盤研究(B) 25280035 により行われた。ここに記して、謝意をあらわす。

参考文献

- [Bizer 09] Bizer, C., Heath, T., and Berners-Lee, T.: Linked Data - The Story So Far, *International Journal on Semantic Web and Information Systems*, Vol. 5, No. 3, pp. 1-22 (2009)
- [Hoffart 13] Hoffart, J., Suchanek, F. M., Berberich, K., and Weikum, G.: YAGO2: A spatially and temporally enhanced knowledge base from Wikipedia, *Artificial Intelligence*, Vol. 194, No. 0, pp. 28 - 61 (2013), [jce:titlejArtificial Intelligence, Wikipedia and Semi-Structured Resourcesj/ce:titlej](#)
- [吉岡 12] 吉岡 真治: Wikipedia を中心とした Linked Open Data に関する一考察, 情報処理学会デジタルドキュメント研究会, 2012-IFAT-107 (2012), IFAT-107-1
- [玉川 10] 玉川 奨, 桜井 慎弥, 手島 拓也, 森田 武史, 和泉 憲明, 山口 高平: 日本語 Wikipedia からの大規模オントロジー学習, 人工知能学会論文誌, Vol. 25, No. 5, pp. 623-636 (2010)
- [藤原 12] 藤原 嵩大, 吉岡 真治: Wikipedia の階層関係を分析するためのカテゴリパターンの提案, 2012 年度人工知能学会全国大会(第 26 回) 論文集 (2012), CD-ROM 2C1-NFC2-4